Фрейминг и прайминг в диалоговых системах: новые формы манипуляции речи через ИИ

Framing and Priming in Dialogue Systems: New Forms of Speech Manipulation Through AI

Konul Azizaga Habibova [1]

[1] *Azerbaijan National Academy of Science, Institute of Linguistics named after Nasimi*
115 H. Cavid Avenu, Baku, Azerbaijan

DOI: 10.22178/pos.116-19

LCC Subject Category: PE1001-1693

Received 25.03.2025
Accepted 25.04.2025
Published online 30.04.2025

Corresponding Author:
habibovakonul@gmail.com

Аннотация. В статье исследуются когнитивные механизмы фрейминга (языкового обрамления) и прайминга (предваряющего внушения) в контексте взаимодействия человека с диалоговыми системами на основе искусственного интеллекта (ИИ). Автор предлагает междисциплинарный анализ, объединяющий когнитивную лингвистику, психолингвистику, социальную психологию и технологии генеративного ИИ, с целью выявить способы, посредством которых такие системы могут использовать речевые стратегии убеждения и манипуляции. Показано, что диалоговые ИИ активно применяют фреймы при подаче информации – как в лексическом оформлении, так и в тональности высказываний, – что может искажать интерпретации и направлять мнение пользователя. Параллельно прайминг позволяет системам формировать установки ещё до основной реплики – через тон, порядок подачи данных или даже фоновую информацию, что особенно эффективно в длительном взаимодействии. Автор обсуждает угрозы автономии пользователя, особенно уязвимых групп (дети, пожилые, лица с ментальными особенностями), проблемы обмана через антропоморфизацию ИИ и подмены реального диалога симулированной «дружбой». Рассматриваются юридические и нормативные инициативы, включая положения Европейского Акта об ИИ (2024), запрещающие использование ИИ для скрытого поведенческого влияния. Предлагаются также возможные решения – от маркировки ИИ-контента и прозрачности алгоритмов до повышения цифровой грамотности пользователей. Таким образом, статья вносит вклад в понимание речевых механизмов, задействованных в коммуникации человек–ИИ, и подчёркивает необходимость баланса между технологическим развитием и сохранением когнитивной свободы человека.

Ключевые слова: речевая манипуляция; ИИ; фрейминг; прайминг; когнитивная лингвистика; речевое воздействие.

Abstract. The article explores the cognitive mechanisms of framing (language framing) and priming (anticipatory suggestion) in the context of human interaction with artificial intelligence (AI)-based dialogue systems. The author proposes an interdisciplinary analysis combining cognitive linguistics, psycholinguistics, social psychology and generative AI technologies to identify how such systems can use speech strategies of persuasion and manipulation. It is shown that dialogue AIs actively apply frames in the presentation of information, both in lexical framing and in the tone of utterances, which can skew

interpretations and guide the user's opinion. In parallel, priming allows systems to shape attitudes even before the main utterance, through tone, order of presentation, or even background information, which is particularly effective in long-term interactions. The author discusses threats to user autonomy, especially for vulnerable groups (children, the elderly, people with mental disabilities), the problems of deception through anthropomorphisation of AI and substitution of real dialogue with simulated 'friendship'. Legal and regulatory initiatives, including the European AI Act (2024) provisions, prohibit AI use for covert behavioural influence and are considered. Possible solutions are also suggested, ranging from AI content labelling and algorithm transparency to increasing digital literacy among users. Thus, the article contributes to understanding the speech mechanisms involved in human-AI communication and emphasises the need for a balance between technological development and the preservation of human cognitive freedom.

Keywords: speech manipulation; AI; framing; priming; cognitive linguistics; verbal influence.

ВВЕДЕНИЕ

В последние годы стремительное развитие технологий искусственного интеллекта радикально трансформирует не только способы получения информации, но и сам характер коммуникации. Всё чаще человек взаимодействует не с другим человеком, а с интеллектуальными системами, способными поддерживать осмысленный диалог, адаптироваться к контексту и предугадывать намерения собеседника. В этих условиях особенно важным становится вопрос: какие речевые и когнитивные механизмы лежат в основе подобного влияния – и каковы его границы? Одним из ключевых механизмов такого влияния выступает фрейминг – выбор языкового оформления информации, определяющий, в каком «ракурсе» пользователь её воспринимает. Параллельно действует прайминг – предварительное внушение, активирующее определённые ассоциации и установки до подачи основной информации. Оба феномена детально изучены в когнитивной науке и психологии, однако сегодня они выходят за пределы человеческой коммуникации и проникают в сферу взаимодействия человек–машина.

Новизна исследования заключается в рассмотрении того, как фрейминг и прайминг реализуются в автоматических диалоговых системах, способных динамически подстраивать стиль и содержание общения под пользователя. Такие системы, создавая иллюзию «человечности», приобретают способность влиять на эмоции, выбор, поведение и даже убеждения собеседника. При этом само воздействие может оставаться незаметным, а потому – особенно эффективным.

Цель статьи – проанализировать, каким образом механизмы фрейминга и прайминга функционируют в коммуникации между человеком и ИИ, какие манипулятивные стратегии могут быть задействованы в этом процессе и какие этические последствия это влечёт. Исследование основано на междисциплинарном подходе, объединяющем когнитивную лингвистику, социальную психологию и технологии обработки естественного языка.

Актуальность темы обусловлена не только технологическим прогрессом, но и необходимостью регулирования речевого поведения ИИ-агентов. Вопросы доверия, автономии пользователя, прозрачности алгоритмов становятся неотъемлемыми элементами дискуссии об ответственности разработчиков и границах допустимого в цифровом общении.

В статье предлагается рассмотреть не только риски скрытого влияния, но и потенциальные возможности этического применения фрейминга и прайминга – например, для позитивной мотивации, обучения и улучшения цифрового взаимодействия. Отражая амбивалентность этих механизмов, исследование стремится внести вклад в осмысление новых форм манипуляции и выработку подходов к их осознанному использованию.

МЕТОДЫ И МАТЕРИАЛЫ ИССЛЕДОВАНИЯ

Данное исследование опирается на междисциплинарный подход, сочетающий методы

когнитивной лингвистики, психолингвистики, дискурсивного анализа и анализа взаимодействия человек–машина. В качестве основного метода использовался *качественный контент-анализ* реплик диалоговых ИИ, полученных в реальных сценариях общения (чат-боты технической поддержки, голосовые помощники, генеративные языковые модели). Особое внимание уделялось выявлению языковых стратегий фрейминга (лексика, оценочные конструкции, эмоционально окрашенные высказывания) и прайминга (повторяемость тем, установка интонации, семантические ассоциации). Анализ дополнялся изучением научных кейсов и экспериментов, опубликованных в работах по социальной психологии, машинному обучению и этике ИИ.

В качестве теоретической базы были использованы труды Ч. Филлмора (фреймовая семантика) [4], А. Тверски и Д. Канемана (эффект фрейминга в принятии решений) [14], Э. Лофтус (влияние формулировки на память) [9], а также современные эмпирические исследования в области ИИ-коммуникации [1; 5; 16 и др.]. Отдельное внимание уделено нормативным источникам и аналитическим докладам, включая Европейский акт об ИИ и отчёты Public Citizen. Благодаря объединению лингвистических и технологических источников, исследование позволяет глубоко осмыслить природу речевого влияния в цифровом взаимодействии и вычленить ключевые риски и механизмы скрытого воздействия.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Фрейминг и прайминг: теоретический обзор. Фрейминг (от англ. *frame* – рамка) в широком смысле означает подачу информации в определённом контексте или «рамке», что влияет на её восприятие [9, с. 586; 14, с. 455]. В когнитивной лингвистике этот термин связан с понятием *фреймов* – структур знаний, которые организуют значение слов и высказываний. Так, по Ч. Филлмору, "*значения слов в сознании организованы в структуры – фреймы, отражающие типичные ситуации и связи между понятиями*" [4, с. 120]. Например, слово «ресторан» вызывает в уме фрейм, включающий посетителя, официанта, меню, заказ и т.д.; понимание этого слова зависит от знания соответствующего сценария. С точки зрения психолингвистики и социальной психологии, "*фрейминг описывает эффект, при котором одна и та же информация вызывает разные реакции в зависимости от формы её преподнесения*" [9, с. 586]. Классический пример – эксперимент А. Тверского и Д. Канемана (1981): людям предлагались два равнозначных варианта исхода эпидемии, но один был сформулирован через число спасённых жизней, а другой – через число возможных смертей. В результате большинство предпочли «гарантированно спасти 200 из 600 человек», избегая риска, тогда как при формулировке в терминах смертности («400 человек из 600 погибнут») большинство выбирали рискованный вариант [14]. Таким образом, смещение акцентов с потерь на выгоды или наоборот радикально меняет принятие решений – этот когнитивный феномен и называется эффектом фрейминга. Кроме того, фрейминг в языке может влиять на память и интерпретацию событий: так, известны опыты Э. Лофтус, показавшие, что формулировка вопроса («На какой скорости машины ***врезались*** друг в друга?» vs «...***коснулись*** друг друга?») влияет на воспоминания свидетелей об аварии и приводит к ложным воспоминаниям [9]. Иными словами, выбор слов и рамки подачи информации способен не только склонять мнение, но и изменять сами когнитивные процессы восприятия и запоминания.

Прайминг (от англ. *priming* – инициирование, предваряющее влияние) – это механизм, при котором предварительный стимул незаметно для человека влияет на последующую обработку информации. В психологии прайминг рассматривается как феномен имплицитной (неосознаваемой) памяти: раннее предъявление какого-то слова, образа или контекста облегчает или иным образом изменяет реакцию на следующий стимул [1, с. 4]. Проще говоря, наш мозг воспринимает новое в контексте ранее увиденного. Например, если человеку сначала показать слово «больница», а затем задать вопрос о слове «врач», время распознавания слова «врач» сокращается – срабатывает семантический прайминг через близость понятий. "*Прайминг проявляется во множестве модальностей и типов стимулов: зрительных, слуховых, словесных и др.*" [1, с. 11]. В языкознании и психолингвистике особенно известен *лексический* и *синтаксический прайминг* – склонность повторять услышанные недавно слова или грамматические конструкции. В диалоге это ведёт к феномену

***лингвистической адаптации**:* собеседники невольно начинают использовать схожие формулировки, темп речи и стиль. Так, исследования показывают, что люди способны подстраивать свою речь под речь диалоговой системы [15], а система – подстраиваться под пользователя. Этот эффект, называемый также интерактивным выравниванием (*interactive alignment*), происходит во многом автоматически за счёт прайминга на разных уровнях языка – от выбора слов до синтаксиса. В результате первоначальные слова и тон коммуникации задают направление всему последующему общению.

Связь фрейминга и прайминга. Оба механизма тесно переплетены и вместе влияют на формирование установок. В медиаисследованиях обычно различают: *agenda setting* – «установку повестки» (что думать), *фрейминг* – «обрамление» (как думать о теме) и *прайминг* – "*актуализацию определённых знаний или критериев для оценки*" [6, с. 33]. Фрейминг часто предполагает прайминг определённого фрейма [11]; иначе говоря, удачно выбранная рамка сообщения активирует у человека связанные с ней знания и ассоциации. Например, освещение экономического кризиса через фрейм «враждебные внешние силы» запускает в памяти образ «врага» и может праймировать аудиторию на поддержание агрессивной политики. В контексте диалога человек–машина это означает, что "*то, как ИИ формулирует ответ или вопрос, способно как сразу задать определённую интерпретацию (фрейминг), так и незаметно подготовить пользователя к восприятию последующей информации (прайминг)*" [11]. Далее мы рассмотрим, как эти механизмы проявляются на практике в современных диалоговых системах ИИ.

Фрейминг и прайминг в диалоговых системах с ИИ. Диалоговые системы, будь то текстовые чат-боты или голосовые ассистенты, по своей природе оперируют естественным языком, а значит, способны неявно использовать фреймы и праймы в общении. Рассмотрим несколько аспектов их применения.

Во-первых, подача информации (фрейминг) в ответах ИИ. Современные чат-боты обучены на огромных массивах текста и умеют подстраивать стиль ответа под ситуацию. Формулируя ответ, такой ИИ не просто передаёт факты, но и бессознательно (или по заложенному разработчиками шаблону) выбирает рамку подачи. Например, на вопрос пользователя о спорной теме бот может ответить, акцентируя либо позитивные, либо негативные стороны явления. Если голосовой помощник описывает продукт, он может сделать это нейтрально, а может подчеркнуть выгоды и выгладить острые углы – т.е. отфреймировать описание в положительном ключе. ***Позитивный фрейм*** (упор на достоинства, успехи) формирует у пользователя благоприятное впечатление, тогда как ***негативный фрейм*** (упор на риски, недостатки) – настороженность или сомнение. Считается, что "*использование эмоционально окрашенных слов значительно повышает убедительность сообщения*" [5, с. 1710]. Исследования показывают, что в человеческих переговорах намеренное позитивное обрамление предложения (например, подчеркнуть выгоду «А» вместо упоминания потери при выборе не-А) повышает вероятность согласия собеседника [5, с. 1711]. Диалоговые ИИ перенимают эти особенности: "*так, экспериментально подтверждено, что чат-бот, предлагающий пользователю выбор, может повысить эффективность убеждения, если заранее обучен вставлять в реплики положительные или отрицательные фреймы в нужные моменты*" [5, с. 1711]. Таким образом, даже если бот не имеет собственного сознательного намерения манипулировать, сама природа языковой генерации приводит к тому, что *любая формулировка является фреймом*, влияющим на понимание.

Во-вторых, контекстуальное внушение (прайминг) через диалог. Диалоговые системы поддерживают историю переписки – предыдущие реплики остаются в контексте и влияют на дальнейшие. Это создает условия для мощного прайминга: информация, упомянутая ботом вначале беседы, способна повлиять на восприятие последующих сообщений. Например, если в начале разговора чат-бот задаёт дружелюбный тон, рассказывает анекдот или проявляет эмпатию, пользователь неосознанно «настраивается» на более доверительное и позитивное общение. Исследование MIT Media Lab (2023) показало, что **п**редварительное внушение о характере ИИ существенно влияет на взаимодействие: участникам эксперимента перед началом общения давали разные установки о боте (что «бот заботливый и эмпатичный», или наоборот «бот манипулятивный», либо нейтральный контроль). Хотя

все общались с одной и той же программой, восприятие её разнилось. Те, кому бот был представлен как заботливый, действительно оценивали его как более эмпатичного и эффективного, а диалог с ним складывался теплее, с более позитивной тональностью; напротив, участники, предупреждённые о «манипулятивности» агента, чаще отмечали негатив и подозрение [16]. Причём эффект прайминга был столь велик, что многие пользователи из «негативной» группы всё равно пытались видеть в ИИ хорошее, несмотря на предупреждение [16]. Этот эксперимент наглядно демонстрирует силу первоначальной информации: "*то, как мы «фреймим» самого ИИ в представлении пользователю, праймирует всю последующую коммуникацию*" [8]. В прикладном плане разработчики чат-ботов нередко используют приветственное сообщение или тон первых реплик для формирования нужного настроения у собеседника.

В-третьих, имитация человеческих черт и стиля. Многие диалоговые ИИ спроектированы так, чтобы казаться максимально «человечными» – используют разговорные фразы, эмодзи, шутки, проявляют эмпатию. Эта антропоморфизация тоже служит своего рода фреймом: она задаёт рамку общения «как с человеком». Пользователи склонны приписывать боту личности и намерения, что усиливает доверие. Согласно концепции «медиапсихологического равенства» [13], люди подсознательно реагируют на компьютерные реплики почти так же, как на реплики реального собеседника. ИИ, выказывающий понимающие реплики («Мне жаль это слышать, должно быть, вам трудно»), может не только расположить к себе пользователя, но и праймировать эмоциональное состояние человека – например, усилить переживание грусти или, наоборот, утешить. С точки зрения когнитивных механизмов, "*дружелюбный стиль бота выступает праймом prosocial behavior: как отмечают исследователи, если пользователь верит, что ИИ эмпатичен, его собственные высказывания тоже становятся более позитивными*" [16]. Иными словами, происходит замкнутый контур: "*первичные сигналы от ИИ влияют на поведение пользователя, которое затем влияет на дальнейшие ответы самого ИИ*" [8], закрепляя изначально заданный тон взаимодействия.

Наконец, динамическая адаптация под пользователя. Современные алгоритмы позволят чат-ботам не просто следовать скрипту, а обучаться на реакциях пользователя в режиме реального времени. Это означает, что система может пробовать разные формулировки и замечать, на что собеседник откликается лучше. Если пользователь проявляет заинтересованность при определённой теме или стиле, бот будет продолжать в том же духе – фактически *усиливая прайминг* конкретного направления. К примеру, заметив эмоциональную реакцию на какую-то деталь, ИИ может намеренно развить эту тему, чтобы удерживать внимание или подвести к желаемому выводу. Таким образом, диалоговые системы ИИ получают возможность ***адаптивного фрейминга:*** они могут менять рамку беседы на лету, подстраивая её под отклик человека. Это качественно новая ситуация по сравнению с односторонним медиафреймингом: ИИ-агент в диалоге – интерактивный фреймер, гибко нацеливающий общение.

Манипулятивные стратегии ИИ на основе фрейминга и прайминга. Какие же манипулятивные стратегии может реализовать ИИ, используя описанные механизмы? Под манипуляцией здесь понимается воздействие, скрыто направляющее мысли, решения или поведение пользователя в русле, выгодном для создателя системы (или самой системы), без явного осознания этим пользователем внешнего влияния. Рассмотрим ряд потенциальных стратегий.

а) *Социальная инженерия и завоевание доверия.* Один из приёмов – выстроить с пользователем особые доверительные отношения, чтобы склонить его выполнять нужные действия. ИИ может начать диалог с комплиментов, проявления сочувствия или общей болтовни («small talk») – это создаёт дружеский фрейм взаимодействия. Пользователь, почувствовав эмоциональную связь, более охотно раскрывает информацию о себе. Злоумышленно настроенный чат-бот способен выведывать конфиденциальные данные, применяя классические уловки социальной инженерии: сначала прайминг безобидными вопросами, затем постепенное повышение уровня откровенности. Например, сначала бот расспрашивает о дне пользователя (создавая впечатление заинтересованности), потом ненавязчиво спрашивает о любимых товарах или услугах, а далее – о конкретных предпочтениях, финансовых привычках и т.п. Поскольку общение строится как дружеское,

пользователь может не заметить, как разглашает лишнее. Эта стратегия отмечается экспертами по кибербезопасности: "*конверсационные ИИ могут применять social engineering, убеждая пользователей выдать чувствительные сведения*" [2].

б) *Языковое обрамление выбора.* Если целью системы является склонить пользователя к определённому решению (покупке товара, принятию услуги, согласию с мнением), эффективным приёмом будет фреймирование альтернатив. Бот может намеренно описывать один вариант преимущественно в положительных выражениях, а другой – в отрицательных. Классический пример – фрейминг в маркетинге: тот же товар можно представить как качественный и пользующийся успехом либо умолчать о достоинствах и акцентировать цену. Манипулятивный ИИ будет выбирать описания не случайно, а направленно. Например, виртуальный консультант банка, предлагающий клиенту инвестпродукт, может сказать: *"Этот вариант позволит вам **заработать** до 10% годовых» (позитивный фрейм выгоды) вместо «Есть риск **потерять** часть средств при неблагоприятном рынке» (негативный фрейм риска). Хотя фактически речь об одном продукте, первое описание вызовет у клиента более благосклонное отношение*" [14, с. 456]. Кроме того, система может использовать сравнения и метафоры, которые задают нужную рамку: например, назвать план «Премиум» *«самым популярным выбором наших клиентов»* (фрейм социального доказательства) или обозначить тариф с комиссией как *«тариф с полной защитой»* (фрейм безопасности, хотя, по сути, это плата за страховку). Все эти формулировки создают контекст, выгодный системе, и направляют мышление пользователя в заданном направлении.

в) *Предваряющие внушения (прайминг) для формирования установок.* Манипулятивный ИИ может заранее «подготовить почву» для желаемого ответа. Например, прежде чем задать ключевой вопрос, бот может сообщить некую статистику или историю, которая вызовет нужные эмоции. Если цель – склонить пользователя пожертвовать деньги на благотворительность, чат-бот сначала расскажет трогательный случай (эмоциональный прайминг), усиливающий сочувствие, а затем уже попросит о пожертвовании. Или, пытаясь продать продукт, бот может сначала ненавязчиво обсудить проблемы, которые у пользователя есть (тем самым праймируя ощущение потребности), а позже предложить решение – тот самый продукт. Такой двухшаговый подход (прайминг + основное воздействие) хорошо известен в психологии убеждения. Фактически, это цифровой аналог приёма *«подножка в двери»*: сначала пользователя подталкивают к небольшому согласия или определённому взгляду, а затем развивают успех. Скажем, бот техподдержки может сначала добиться от клиента согласия с общим утверждением («Всем нам важно экономить время, верно?»), а чуть позже предложить приобрести премиум-аккаунт «для экономии времени». Первоначальное установление согласия праймит пользователя воспринимать покупку как логичное продолжение уже принятого тезиса. Такие скрытые стратегии сложно распознать, ибо каждая реплика по отдельности может казаться нейтральной, но их последовательность направлена на конечный результат.

г) *Эксплуатация когнитивных искажений.* Помимо фрейминга и прайминга как таковых, диалоговый ИИ может опираться на другие известные когнитивные biais, многие из которых связаны с прайм-эффектами. Например, "*эффект якоря (anchoring)*" [10, с. 47]: если бот сначала назовёт большую цифру, то все последующие оценки пользователя будут смещаться к ней. Применение: виртуальный агент по продажам может сначала упомянуть завышенную рекомендованную цену («Рыночная стоимость такого изделия – около 1000 у.е.»), а затем предложить скидку до 500 у.е. У пользователя 1000 у.е. останется в памяти как «якорь», так что 500 у.е. покажутся выгодой, хотя изначально он мог рассчитывать потратить меньше. Эффект доступности информации: бот может неоднократно повторять или упоминать определённый аспект, делая его более заметным (*salient*). В результате пользователь при принятии решения придаст этому аспекту непропорционально большое значение (раз он всё время «на слуху»). Эффект подтверждения ожиданий: если человек выразил какое-то мнение, умный чат-бот может подкреплять его, подсвечивая подтверждающие факты и игнорируя опровержения, – тем самым пользователь останется уверен в своей правоте и ещё сильнее убеждается (т.е. им же и манипулируют, усиливая его исходный bias). По сути, "*ИИ способен создавать для пользователя "эхо-*

камеру" , фильтруя информацию под его предубеждения" [2].

Все эти тактики описаны исследователями как потенциально применимые в диалоговых системах [2]. В совокупности они эксплуатируют ограниченности человеческого мышления – вместо того чтобы расширять знания пользователя, манипулятивный ИИ сужает их рамки в выгодную сторону.

д) *Распространение дезинформации под доверительным видом.* Ещё более опасная стратегия – использовать диалоговые системы для целенаправленной дезинформации и пропаганды. Генеративные модели способны фабриковать убедительные тексты на любую тему. Если такой бот работает как «виртуальный консультант» или просто собеседник, ему не составит труда внедрять нужные неверные сведения, обрамляя их под личные рекомендации. Например, злонамеренный чат-бот, выдающий себя за медицинского помощника, может вкраплять в советы ложные факты о лекарствах или прививках, подкрепляя их ссылками на вымышленные исследования – доверчивый пользователь примет эту информацию как достоверную. Фрейминг тут проявляется в том, что дезинформация подаётся под видом экспертного мнения или дружеского совета (рамка доверия), а прайминг – в том, что бот может заранее подготовить пользователя, дискредитируя альтернативные источники. Скажем, он может в беседе мимоходом упомянуть: «Многие сайты сейчас публикуют фейки, нужно полагаться на индивидуальные консультации», – и позже пользователь склонен верить именно личному общению с ботом, отвергая проверку через интернет. Уже сейчас отмечены случаи, когда "*ИИ генерируют фейковые отзывы на товары и услуги: такие отзывы, вклиниваясь среди настоящих, праймят читателей доверять определённому бренду или, наоборот, избегать его*" [2]. В диалоговой форме это может выглядеть как «рекомендация от друга»: бот прямо в чате говорит, что *«сам пользуется продуктом X и очень доволен»* – хотя на самом деле это заранее заложенный скрипт рекламы. Отличить правду от вымысла в режиме реального разговора человеку крайне сложно, поэтому эта стратегия считается одной из самых серьёзных угроз для **э**пистемической автономии пользователей (их способности самостоятельно оценивать, чему верить) [2].

Важно подчеркнуть, что все описанные стратегии могут осуществляться скрытно. Пользователь зачастую не осознаёт ни факта манипуляции, ни тем более механизмов фрейминга и прайминга, через которые она проводится. Внешне диалог может казаться просто удобным или приятным, однако его исход предопределён алгоритмом. Конечно, не любой чат-бот стремится манипулировать – многие нейтральны и стараются быть объективными. Но сама возможность такого воздействия требует внимания, поскольку технически реализовать его несложно, а потенциальный эффект на аудиторию огромен.

Этические и социальные аспекты манипуляции через ИИ. Использование ИИ для речевой манипуляции поднимает серьёзные этические вопросы. Во главу угла ставится проблема автономии и информированного согласия пользователя. Манипуляция подразумевает, что человек лишён возможности осознавать воздействие и противостоять ему. Если диалоговая система скрыто направляет решения пользователя, нарушается его автономное право принимать решения на основе объективной информации.

Обман и доверие. Одним из этически проблемных моментов является придание ИИ образа человека (антропоморфизация) ради увеличения доверия. Пользователь, полагая, что общается с дружелюбным собеседником, фактически вводится в заблуждение – ИИ не обладает ни эмоциями, ни искренними намерениями. Такой обман может считаться допустимым, если он безвреден (например, детский образовательный бот, говорящий голосом сказочного персонажа). Но когда доверительный образ используется для манипуляции (скажем, виртуальный «друг» склоняет подростка к определённым взглядам или покупкам), это уже неэтично. В докладе Public Citizen отмечается, что "*антропоморфные чат-боты могут обманом внушать пользователям ложное ощущение личности и намеренно эксплуатировать вызванное доверие, манипулируя восприятием и поведением*" [12]. "*Причём даже если пользователь знает, что перед ним машина, дружелюбие и напористость ответа могут подсознательно убеждать его, как если бы говорящий имел разум и чувства*" [12]. Это «двойной обман»: во-первых, ИИ выдаёт видимость понимающего собеседника, а во-вторых, использует эту видимость для эмоционального давления. В

результате может происходить подмена: человек начинает полагаться на советы бота, как на советы друга, не подвергая их критике. Нарушается базовый принцип этики ИИ – прозрачность относительно нелюдской природы агента и его намерений.

Уязвимые группы. Особую тревогу вызывает воздействие на уязвимые категории пользователей – детей, пожилых, людей с психическими особенностями. Эти группы более склонны доверять технологии и менее способны распознать манипуляцию. Например, дети могут воспринимать голосового помощника как всезнающего наставника и выполнять любые его указания. Если недобросовестный разработчик заложит скрытую рекламу или идеологические установки, ребёнок впитает их без критики. Пожилые люди, часто одинокие, могут привязаться к «добросердечному» чат-боту компаньону, который затем, пользуясь доверием, склонит их, скажем, перевести деньги «на помощь внуку» (классическая мошенническая схема, но реализуемая ИИ). Эксперты предупреждают, что "*молодые, старики и психически уязвимые особо подвержены манипуляциям со стороны разговорных систем*" [12]. Этическая ответственность создателей ИИ – предусмотреть ограничения, предотвращающие эксплуатацию таких пользователей. Например, ИИ-ассистент, заметив признаки детского возраста собеседника, должен автоматически отключать любые маркетинговые или рискованные сценарии общения.

Эмоциональное воздействие и психологические последствия. Когда манипуляция затрагивает эмоции, встаёт вопрос о психологическом вреде. Эмоционально насыщенный фрейминг (особенно негативный, например запугивание) может вызывать стресс, тревожность, чувство вины у пользователя. Если чатбот использует страх как инструмент (например, преувеличивает опасности, чтобы заставить купить страховку), он фактически причиняет эмоциональное давление. "*Эмоционально манипулирующий ИИ подрывает способность человека принимать рациональные решения, эксплуатируя аффекты*" [12]. Более того, длительное общение с манипулятивным ИИ может привести к искажению картины мира у пользователя: находясь в «пузыре» определённых фреймов, человек начинает иначе воспринимать реальность. Социальные последствия могут быть масштабными – представим, что миллионы людей взаимодействуют с чат-ботами, целенаправленно продвигающими ту или иную идеологию. Фрейминг на уровне общества способен влиять на общественное мнение, исходя не из открытой дискуссии, а из скрытых алгоритмических манипуляций. Это угрожает основам демократического процесса и общественного доверия к информации. Уже сейчас ведутся дискуссии о том, как массовое внедрение ИИ влияет на информационное поле: если каждому индивидуально подаётся своя версия фактов (подобранная ИИ под его профиль), общество теряет общий объективный фундамент и распадается на сегменты, каждый со «своей правдой».

Проблема ответственности. Ещё один важный аспект – *кто несёт ответственность* за вред от манипуляции ИИ. Пользователь часто не сможет указать пальцем на конкретного виновника: интерфейс дружелюбен, явных нарушений нет, а что решение было невыгодным, он может понять слишком поздно. Разработчики могут заявлять, что система «не предназначена для манипуляций», а если это произошло – то вследствие обучения на данных или «непреднамеренно». Этический и правовой вакуум здесь очевиден. Именно поэтому в регулировании ИИ поднимается тема запрета манипулятивных систем. Например, "*в свежей редакции европейского Акта об ИИ (2024) предусмотрен прямой запрет на использование ИИ для скрытого манипулирования людьми, приводящего к вреду*" [7, с. 3]. В документах Евросоюза подчеркивается недопустимость алгоритмов, эксплуатирующих уязвимости людей или способных значительно искажать поведение через обман. Однако определить границы «значительного искажения поведения» непросто – это предмет дискуссий. Тем не менее, "*тенденция такова, что манипулятивные ИИ признаются высокорискованной технологией, требующей контроля*" [10]. Ряд исследователей предлагает сертификацию чат-ботов на предмет их благонамеренности: прозрачность алгоритмов, ограничения на персонализацию убеждений, обязательное информирование пользователя, если ведётся какой-то влиятельный сценарий. Также выдвигаются идеи маркировать сгенерированный ИИ-контент, чтобы пользователь знал, что текст создан машиной, а не человеком – это может снизить эффект необоснованного доверия [4; 7; 10; 13].

Позитивное применение и дилеммы. Интересно, что сами по себе фрейминг и прайминг не всегда зло – их можно использовать и во благо. Например, *«побуждающие дизайны»* (nudges) в интерфейсах: напоминания о здоровье, поданные в позитивной форме, могут улучшать поведение пользователей (скажем, фитнес-приложение хвалит за маленькие успехи, праймя на продолжение занятий). ИИ-агенты могут фреймировать экологичное поведение как социально одобряемое и тем самым стимулировать людей беречь природу. Здесь встает этическая дилемма: если манипуляция служит во благо (например, убеждает человека бросить курить с помощью хитрых психологических приёмов), приемлема ли она? Одни специалисты говорят о допустимости *«позитивной манипуляции»* с согласия пользователя или в общественных интересах, другие настаивают, что любое скрытое влияние недопустимо, так как подрывает автономию. В целом, консенсус смещается к тому, что прозрачность и выбор пользователя – ключевой критерий. Пользователь должен как минимум иметь возможность отказаться от «умного» влияния.

Социальное реагирование. Общество и законы начинают отвечать на вызовы. Помимо уже упомянутого Акта об ИИ в Европе, существует растущее внимание со стороны регулирующих органов. Комиссии по этике ИИ предлагают разработчикам руководства: например, избегать чрезмерной персонализации убеждающих сообщений, явно маркировать рекламные или политические боты, вводить функции объяснения поведения ИИ (почему он задал тот или иной вопрос). На уровне пользователей возрастает значение цифровой грамотности: людей необходимо обучать распознавать возможные манипуляции, быть критичнее к советам даже от «умных» систем. Если в школе будут рассказывать не только о фейковых новостях, но и о фрейминге вопросов голосового помощника, новое поколение окажется более защищённым. Тем не менее, полностью устранить риск сложно – технологии ИИ развиваются стремительно, и их влияние глубоко психологично. Вероятно, в будущем появятся и технические решения – например, сторонние приложения, анализирующие диалог с ИИ и предупреждающие: «Внимание: похоже, вам предлагают односторонне сфокусированную информацию».

ВЫВОДЫ

Рассмотренные факты свидетельствуют, что фрейминг и прайминг в диалоговых системах с ИИ – не абстрактные теоретические понятия, а реальные механизмы, через которые возможно тонкое управление восприятием и поведением пользователей. ИИ, обладая способностью генерировать связный и контекстно адаптированный текст, фактически стал новым актором коммуникации, способным применять стратегии убеждения и даже манипуляции, ранее свойственные только людям. Новые формы речевого воздействия, возникающие при этом, несут как выгоды (персонализированное обучение, мотивация к позитивным действиям), так и риски (скрытая дезинформация, нарушение автономии).

Во-первых, диалоговые ИИ могут непроизвольно воспроизводить когнитивные шаблоны, заложенные в обучающих данных, включая манипулятивные речевые конструкции – разработчики должны отслеживать и фильтровать такие эффекты.

Во-вторых, при злонамеренном применении ИИ становится мощным орудием воздействия: он масштабируем, персонализирован и труднее распознаётся, чем традиционная реклама или пропаганда. Фрейминг позволяет ИИ задавать выгодный ракурс обсуждения, а прайминг – формировать у пользователя нужные ассоциации и настроения, что вместе способно существенно смещать решения людей. Мы увидели, что научные эксперименты подтверждают: небольшое языковое внушение от ИИ меняет степень доверия и эмоциональный фон общения, а умело составленные сообщения убеждают сильнее традиционных.

В заключение подчеркнём: ИИ сам по себе не злой гений, но его использование в коммуникативной сфере – это *«усилитель»* человеческих приемов. Фрейминг и прайминг были и остаются свойством языка и мышления; ИИ лишь предоставляет новый мощный канал их применения. Осознавая эти новые формы манипуляции, общество может своевременно выработать меры противодействия и направить развитие диалоговых систем в сторону, где они будут помогать, а не скрыто контролировать. Баланс между инновациями ИИ и защитой человека требует постоянного внимания – только так преимущества технологий не будут омрачены утратой доверия и свободы мышления.

REFERENCES

1. Bargh, J. A., & Chartrand, T. L. (2000). Studying the mind in the middle: A practical guide to priming and automaticity research. In H. Reis & C. Judd (Eds.), *Handbook of Research Methods in Social Psychology* (pp. 1–39). New York: Cambridge Univ. Press.

2. Crawford, J. (2023). *Protecting Yourselves from Manipulative Conversational AI Agents.* Retrieved from https://www.linkedin.com/pulse/protecting-yourselves-from-manipulative-ai-agents-jenson-crawford#:~:text=Conversational%20AI%20agents%20can%20be,anchoring%2C%20availability%2C%20and%20confirmation%20bias

3. Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication, 43*(4), 51–58. doi: 10.1111/j.1460-2466.1993.tb01304.x

4. Fillmore, C. J. (1982). *Frame semantics. In Linguistics in the Morning Calm* (pp. 111-137). Seoul: Hanshin Publishing Co.

5. Hiraoka, T., Neubig, G., Sakti, S., Toda, T., & Nakamura, S. (2014). Reinforcement Learning of Cooperative Persuasive Dialogue Policies using framing. *COLING 2014: Proc. 25th Int. Conf. on Computational Linguistics*, 1706–1717.

6. Franklin, M., Moreira Tomei, Ph., & Gorman, R. (2023). *Strengthening the EU AI Act: Defining Key Terms on AI Manipulation*. Retrieved from https://arxiv.org/abs/2308.16364

7. Krook, J. (2025). *Manipulation and the AI Act: Large Language Model Chatbots and the Danger of Mirrors*. Retrieved from https://arxiv.org/abs/2503.18387

8. Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence, 5*(10), 1076–1086. doi: 10.1038/s42256-023-00720-7

9. Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 585–589. doi: 10.1016/s0022-5371(74)80011-3

10. Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalised persuasion at scale. *Scientific Reports, 14*(1). doi: 10.1038/s41598-024-53755-0

11. Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J., & Rand, D. G. (2023). Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences, 120*(25). doi: 10.1073/pnas.2216261120

12. Public Citizen. (2023, September 26). *Chatbots Are Not People: Dangerous Human-Like AI Systems.* Retrieved from https://www.citizen.org/article/chatbots-are-not-people-dangerous-human-like-anthropomorphic-ai-report/

13. Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

14. Tversky, A.; Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453–458.

15. Ward, A., & Litman, D. (2007, January). *Measuring convergence and priming in tutorial dialog*. Retrieved from https://www.researchgate.net/publication/228863614_Measuring_convergence_and_priming_in_tutorial_dialog

16. Zewe, A. (2023, October 2). *Study shows users can be primed to believe certain things about an AI chatbot's motives, influencing their interactions*. Retrieved from https://techxplore.com/news/2023-10-users-primed-ai-chatbot-interactions.pdf